DOCUMENT RESUME

ED 432 609 TM 029 992

AUTHOR	Arenson, Ethan A.
TITLE	Statistical Linkages between State Education Assessments and the National Assessment of Educational Progress.
PUB DATE	1999-03-31
NOTE	25p.; Paper presented at the Annual Meeting of the Sacramento Statistical Association (Sacramento, CA, March 31, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; Elementary Education; Grade 4; Grade 8; National Competency Tests; Reading Tests; Sampling; Standardized Tests; *State Programs; *Statistical Analysis; *Test Results; *Testing Programs
IDENTIFIERS	*Linkage; *National Assessment of Educational Progress; State Reading Assessments

ABSTRACT

The National Assessment of Educational Progress (NAEP) measures the educational achievement of nationally representative samples of students in grades 4, 8, and 12. Local educational agencies tend to view the NAEP as a benchmark to which the educational achievement of their students can be compared. In particular, state departments of education wish to compare their assessments to the NAEP. The complex design of the NAEP renders simple comparisons problematic at best. A linear projection-plus-variation method is used to translate student state assessment scores onto the NAEP scale. The accuracy of this method is estimated through repeated half-sampling. A brief description of the 1998 NAEP reading assessment for grades 4 and 8 is discussed. Results from one of the six states in the current linkage study and recommendations for establishing a successful linkage are also presented. (Contains 2 tables and 12 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED 432 609

Statistical Linkages Between State Education Assessments and the National Assessment of Educational Progress

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Ethan Arenson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Ethan A. Arenson
John C. Flanagan Research Center
American Institutes for Research
Palo Alto, California

TM029992

Paper presented at the Annual Meeting of the Sacramento Statistical Association, Sacramento, CA, March 31, 1999.

BEST COPY AVAILABLE

ABSTRACT

The National Assessment of Educational Progress (NAEP) measures the educational achievement of nationally representative samples of students in grades four, eight and twelve. Local educational agencies tend to view NAEP as a benchmark to which the educational achievement of their students can be compared. In particular, state departments of education wish to compare their assessments to NAEP. The complex design of NAEP renders simple comparisons problematic at best. A linear projection-plus-variation method is used to translate student state assessment scores onto the NAEP scale. The accuracy of this method is estimated through repeated half-sampling. A brief description of the 1998 NAEP reading assessment for grades four and eight is discussed. Results from one of the six states in the current linkage study and recommendations for establishing a successful linkage are also presented.

In Fall 1998, six states agreed to participate in a linkage study through which their state assessment results could be used to predict scores on the National Assessment of Educational Progress (NAEP). This paper: (1) provides a brief background on NAEP, (2) discusses the methodology used in constructing the linkage, and (3) identifies the requirements for a successful linkage. The methodology proposed in this paper is justified on the grounds that it seems to work, and not from theoretical considerations.

The Issue of Linking

The need to link state assessments to NAEP arises from political considerations. As NAEP continues to be used to measure achievement on a national or state level, local agencies (typically state departments of education, but possibly county offices of education or even school districts) desire to compare their assessment results to those from NAEP. Ideally, these agencies would administer their assessments, then apply the linkage methodology to determine the achievement level for each student.

There are several reasons for which the issue of linking is problematic. The first involves the estimation of individual scores. In state assessments, the objective typically is to estimate individual performance. States accomplish this objective by administering the same assessment to all students being assessed. However, NAEP is not designed to measure individual proficiency.

Another problem with linking between the two assessments is content-related. NAGB determines the content of NAEP. Consequently, the content of NAEP may not agree with the

content of the state assessments. Psychometrically speaking, one could argue that NAEP and the assessment of a given state do not measure exactly the same construct. They do, however, measure related constructs.

What is the National Assessment of Educational Progress?

Since 1970, NAEP has measured the educational achievement of young Americans in reading, writing, mathematics, science, U.S. history, and geography. Often referred to as "The Nation's Report Card," it accomplishes this task by collecting information on nationally representative samples of students who were either: (1) 9, 13 and 17 years old; or (2) in grades 4, 8 and 12 (Allen & Johnson, 1996).

In each of the subject areas, NAEP estimates the achievement of a *group* of students as a composite score, based on a weighted average of subscales. Typical groups that NAEP uses for reporting include gender, ethnicity, urbanicity (e.g., students from rural or urban areas), state and nationwide. The distribution of scores are used to determine cut-points for achievement levels (basic, proficient and advanced). The National Assessment Governing Board (NAGB), the entity that oversees NAEP, emphasizes that achievement levels, not scores, are the primary way of reporting NAEP results. (Allen, Johnson, Mislevy, & Thomas, 1996). Through the setting of achievement levels, NAGB can identify what students should know and should be able to do at various points on the NAEP scale.

NAEP Sampling

The point that NAEP estimates achievement for groups of students cannot be overemphasized. For the reading assessment, NAEP employs a matrix-sampling technique, known as partially balanced incomplete-block (PBIB) spiraling[1]. Through PBIB spiraling, each student receives a booklet that contains common blocks, consisting of background and motivational questions, in addition to other blocks that comprise the cognitive items. In the 1998 fourth- and eighth-grade reading assessments, each booklet contained two cognitive blocks, with the exception of one booklet in grade eight that consisted of one "extended" block. Tables 1 and 2 show how blocks were assigned to booklets for grades four and eight.

Estimating Proficiency: Scaling the Items

What follows is a brief description of how NAEP estimates proficiency; a full elaboration can be found in Allen, Johnson, Mislevy, and Thomas (1996). Reported NAEP scores are a weighted composite of subscales. In the 1998 fourth- and eighth-grade reading assessments, the subscales were "reading for literacy" and "reading for information." In addition, NAEP used another subscale for grade eight: "reading to perform a task." For the 1998 assessment, NAEP assigned weights of 0.55 and 0.45 for the literacy and information subscales in grade four. For grade eight, weights of 0.4 were assigned to both the literacy and information subscales, with a weight of 0.2 assigned to the task subscale. Table 3 shows the assignment of blocks to subscales

[1] In the other content areas, NAEP uses a balanced-incomplete-block (BIB) design. The difference between PBIB and BIB designs is that the former consists of sampling with all possible combinations of blocks *within each subscale*.

(N. L. Allen, personal communication, March 15, 1999). Subscale scores were determined by the item responses to specific blocks.

For each subscore, NAEP uses item-response theory (IRT; e.g., Lord, 1980) to generate "plausible values" of proficiency, given student responses to the background and cognitive items. That is, for each of the subscales, an IRT model expresses student tendencies to provide certain responses (e.g., correct or incorrect) to cognitive items as a function of an unobservable parameter, namely achievement.

The questions in the cognitive blocks follow one of four formats: multiple-choice with four foils, short constructed -response, extended constructed-response or open-ended. Tables 4 and 5 outline the item types within each block. Multiple-choice items are dichotomously scored, and scaled with a three-parameter logistic (3PL) model:

$$P(x_j = 1 | \theta_k, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-1.7 a_j (\theta_k - b_j)]}, \qquad (1)$$

where

x_{ij} is the response to item j: 1 if correct; 0, otherwise;

θ_k is the student's proficiency on subscale k;

a_j is the slope parameter, or sensitivity to proficiency, of item j;

b_j is the threshold parameter, or level of difficulty, for item j; and

c_j is the lower asymptote, or chance, parameter associated with students of low proficiency.

Two-point constructed-response and open-ended questions were dichotomously scored. These items were scaled with a two-parameter logistic (2PL) model, which is Equation (1) with the restriction that $c_j = 0$:

$$P(x_j = 1 \mid \theta_k, a_j, b_j) = \left(1 + \exp[-1.7a_j(\theta_k - b_j)]\right)^{-1} \quad (2)$$

Items that were polytomously scored (i.e., scored on a three- or four-point scale) were scaled using a generalized partial-credit model (Muraki, 1992). The generalized partial-credit model is an extension of the 2PL model. For item j with m_j categories, the probability that a student with proficiency θ_k will give a response x_j that falls into the ith category is:

$$P\left[x_j = i \mid \theta_k, a_j, b_j, d_{j,1}, \ldots, d_{j,(m_j-1)}\right] = \frac{\exp\left(\sum_{v=0}^{i} 1.7a_j(\theta_k - b_j + d_{j,v})\right)}{\sum_{g=0}^{m_j-1} \exp\left(\sum_{v=0}^{g} 1.7a_j(\theta_k - b_j + d_{j,v})\right)}, \quad (3)$$

where, θ_k, a_j and b_j are previously defined in Equation (1), and

$d_{j,1}, \ldots, d_{j,m_j-1}$ denote the threshold, or difficulty, parameter of responding in the ith category.

Estimating Proficiency: Plausible Values

It was stated before that one of the purposes of NAEP is to permit inferences about the proficiency of groups of students. If the proficiencies of individual students were known, then standard statistical techniques would be appropriate for making such inferences. Given that individual proficiencies are not known, however, NAEP estimates group proficiency by using students' plausible values of proficiency. Plausible values are scores randomly drawn from the student's "proficiency distribution," and incorporate significant variation to reflect the error due to sampling students with so few questions (Mislevy, 1991). For each student, NAEP generates five plausible values per subscale. It is important to note that plausible values for a given student

have no interpretation, other than serving as an intermediate step in determining the proficiency of groups.

Following Rubin's (1987) method of multiple imputations, student proficiencies are considered to be "missing data." Group statistics (e.g., a group mean, regression coefficient, percent of students at or above each achievement level, etc.) can be approximated by its expectation, given the observed pattern of responses to the cognitive and background items. If $\boldsymbol{\theta}$ represents the vector of student proficiencies, and $\mathbf{x}$ and $\mathbf{y}$ represent, respectively, the matrices of student responses to the cognitive and background items, then the parameter $t(\boldsymbol{\theta},\mathbf{y})$ can be estimated by the statistic

$$t^*(\mathbf{x},\mathbf{y}) = E\left[t(\boldsymbol{\theta},\mathbf{y}) \mid \mathbf{x},\mathbf{y}\right] = \int t(\boldsymbol{\theta},\mathbf{y})\, p(\boldsymbol{\theta} \mid \mathbf{x},\mathbf{y})\, d\boldsymbol{\theta} \ . \qquad (4)$$

NAEP computes the variability of the estimates by jackknife procedures.

NAEP calculates proficiency distributions by comparing, for each student in the group, the mean plausible value with the cutpoints for that grade level. The cutpoints for the basic, proficient and advanced proficiency levels on the 1998 reading assessment were the same as those used in the 1994 reading assessment: For fourth grade, the cutpoints were 208, 238 and 268, respectively, and for grade eight 243, 281 and 323. Table 4 shows the average scaled score (for public schools) and percentage of students at or above each proficiency level for the nation (reported values), as well as the approximate averages for the state in the study (Ballator, Jerry, & Rogers, 1999).

Methodology

Sampling

The two-stage sampling design employed was implemented by Westat, Inc., the sampling subcontractor for NAEP. In this design, the primary sampling unit was the school, and within each school students were randomly selected to participate. Under-represented groups, such as Black or Hispanic students, were oversampled in order to increase the reliability of estimates for these groups of students. Wallace & Rust (1996) provide detailed information about the sampling methodology. For the state presented in this study approximately 3,000 fourth grade students in 100 schools and 3,000 eighth grade students in 100 schools were sampled for the 1998 NAEP main reading assessment.[2] Of these, roughly 2,500 students in each grade produced data from which plausible values could be estimated. The students used for the linkage study were those who participated both in the state assessment and in NAEP.

The Linkage Model

Discussions about linking assessments typically include one or more of the following methodologies: equating, calibration, and projection (Linn, 1993). Equating is the most rigorous of the methods, permitting the comparison of two tests that are similar in content but have different levels of difficulty (Kolen & Brennan, 1995). Tests that are being equated typically, though not always, have items in common. Because the assumption of similar content between NAEP and the state assessments cannot be guaranteed, however, any linkage based on equating methods is suspect.

---

[2] There are two types of NAEP assessments in each subject area: the "main" assessment and the "long-term trend" assessment. Sampling for the main and long-term trend assessments are conducted separately. Because of logistical considerations, sampled schools participate exclusively in either the main assessment or the long-term-trend assessment.

Calibration, a less rigorous procedure, is based on IRT models, and typically requires that the two assessments measure the same construct (Williams, Rosa, McLeod, Thissen & Sanford, 1999). Furthermore, IRT methods use item-level data; states, or other agencies interested in establishing a linkage, most likely will not have item responses for both the state assessment and NAEP. As such, a linkage based on item responses is not appropriate.

The approach used in this study involves projecting state assessment scores onto the NAEP scale, and using these projected scores to estimate proficiency, in terms of the "basic," "proficient," and "advanced" achievement levels. Again, it must be emphasized that it is not appropriate for one to draw inferences from individual projected NAEP scores, due to the large amount of variability that is inherent in the projection. Rather, the individual projections merely serve as an intermediate step in calculating the group percentages at each of the achievement levels.

Calculating Projected NAEP Scores

In essence, the projection is a prediction of the expected NAEP score, given (1) a state assessment score and (2) the linear regression of NAEP scores on state assessment scores for students in the linkage sample. In the simplest case, this projection could be represented as

$$y_{ij} = \beta_0 + \beta_1\left(x_{ij} - \bar{x}_{i\cdot}\right) + \beta_2\bar{x}_{i\cdot} + \varepsilon_{ij}, \qquad (4)$$

where for the *j*th student in the school *i*, y_{ij} denotes the mean of the five NAEP plausible values, x_{ij} the student's state assessment score, $\bar{x}_{i\cdot}$ the mean state assessment score for the *i*th school, and

ε_{ij} random error. The second term is proportional to how well a particular student performs on the state assessment, relative to that student's school mean.

For the state presented in this study, two subtests within the state reading assessment were used as predictors of NAEP proficiency. The first subtest (Test A) measured general reading ability, and consisted entirely of items that were dichotomously scored. The second subtest (Test B) measured reading comprehension, and consisted of dichotomous and polytomous items. Because two state assessments were used as predictors, Equation (4) needs to be modified, resulting in Equation 5:

$$y_{ij} = \beta_0 + \beta_1\left(x_{ij,1} - \bar{x}_{i\cdot,1}\right) + \beta_2\bar{x}_{i\cdot,1} + \beta_3\left(x_{ij,2} - \bar{x}_{i\cdot,2}\right) + \beta_4\bar{x}_{i\cdot,2} + \varepsilon_{ij}, \qquad (5)$$

where

$x_{ij,1}$ denotes the score on the first subtest for student j in school i;

$x_{ij,2}$ denotes the score on the second subtest for student j in school i;

$\bar{x}_{i\cdot,1}$ denotes the mean score on the first subtest for school i; and

$\bar{x}_{i\cdot,2}$ denotes the mean score on the second subtest for school i.

In previous linkage studies, certain demographic variables have been known to make the model neutral. Thus, where $d_{ij,1}$, $d_{ij,2}$, $\bar{d}_{i\cdot,1}$ and $\bar{d}_{i\cdot,2}$ denote, respectively, the ethnicity (0 if "white"; 1, otherwise) and gender (0 if male; 1, otherwise) of student j at school i, the proportion

of students at school *i* of ethnicity $d_{ij,1}$, and the proportion of students at school *i* of gender $d_{ij,2}$, Equation (5) becomes

$$y_{ij} = \beta_0 + \beta_1\left(x_{ij,1} - \bar{x}_{i\cdot,1}\right) + \beta_2\bar{x}_{i\cdot,1} + \beta_3\left(x_{ij,2} - \bar{x}_{i\cdot,2}\right) + \beta_4\bar{x}_{i\cdot,2} + \beta_5\left(d_{ij,1} - \bar{d}_{i\cdot,1}\right) + \beta_6\left(d_{ij,2} - \bar{d}_{i\cdot,2}\right) + \varepsilon_{ij} \quad (8)$$

Equation (8) is referred to as the full model. A simpler model can be used for projection if some coefficients are small with respect to their standard error.

Determining the Variation Distribution

Projection, alone, would be a sufficient method if the goal were to estimate group mean NAEP scores. However, the goal of the linkage is that of estimating the proficiency distribution (i.e., the percent of students at or above the basic, proficient and advanced levels). Used alone, projection tends to underestimate the percentages of students who perform in the tails of the distribution of predicted NAEP scores, and will lead to grossly inaccurate estimates of students performing at either the "below basic" or "advanced" levels. For this reason, variation must be included in the projection to ensure that the variance of the predicted scores matches the variance of the NAEP scores in the linkage sample.

The variation, ε_{ij}, that is added to each projected score follows an empirical error distribution. The error distribution can be considered as the sum of two components, and is assumed to be homogeneous across students. The first of these components is the variability

inherent in the distribution of plausible values for each student, and is the residual from the mean plausible value for each student. For example, a fourth-grade student with a true mean plausible value of 238 (the cutpoint for the proficient level) would be expected to perform at or above the proficient level half of the time.

The second component reflects the lack of perfect correlation between the projected NAEP score and the mean of the NAEP plausible values. If the correlation between the projected and actual NAEP performance is r, then $100(1 - r^2)$ percent of the actual NAEP variance needs to be added to the variation distribution. This component of the distribution may not follow any well-known distribution, and the differences in the tails can be quite extreme. Thus, this component is determined empirically, by computing the distribution of the residuals, $y - \hat{y}$.

Determining the Coefficients

An ordinary least squares regression method is used to determine the coefficients to Equation (8). Starting from the full model, Equation (8), certain factors can be dropped, if their coefficients are found not to be significantly different from zero. Standard errors are determined through half-sample replication. In one replication, half of the schools were randomly assigned to an estimation sample, with the other half assigned to a validation sample. For the estimation sample, the coefficients to the model are estimated based on the regression method mentioned above. Then the standard errors of the estimated coefficients are computed by applying the regression model to the validation sample. The standard errors of the coefficients are taken to be the mean of the errors (of the standard error estimates of the coefficients) in the 100 replications. The doubling of variances due to half-sampling balances the halving of the variance, since the

repeated half-samples come from the same finite universe of schools (McLaughlin & Arenson, 1999).

Estimating Proficiencies for Groups

The percentages of students achieving each of the NAEP proficiency levels are computed for the validation sample, using the first of the five plausible values as the student's hypothetical NAEP score. An achievement level is determined for each student based on the hypothetical score, and is compared to the proficiency levels predicted by the model.

The standard error of the linkage to estimate NAEP performance for a group of students who did not participate in NAEP has two components, and is a function of the size of the size of the sample to which the linkage is to be applied. If n_{stu} and n_{sch} denote the numbers of students and schools in the target sample, and if p_b and p_w denote the proportions of variance in the linkage sample between- and within-schools, n_{eff}, can be computed as follows:

$$n_{eff} = \left(\frac{p_b}{n_{stu}} + \frac{p_w}{n_{sch}} \right)^{-1} \quad (9)$$

For the state in this study, $p_b = 0.9$ and $p_w = 0.1$.

Having computed the effective sample size, the standard error of the projected mean NAEP score for the target sample is given by

$$\hat{\sigma}_p = \sqrt{\hat{\sigma}^2{}_s + \frac{\hat{\sigma}^2{}_t}{n_{eff}}}, \text{ where} \quad (10)$$

$\hat{\sigma}_s$ denotes the estimated variance of the predicted mean NAEP scores due to sampling error; and

$\hat{\sigma}_t$ denotes the variance due to linkage error in the target sample.

Typically, $\hat{\sigma}_s$ ranges from one to three points on the NAEP scale, whereas $\hat{\sigma}_t$ is around 20 points on the NAEP scale.

Results

The mean state scores for the fourth graders in the linkage sample were 40 for Test A and 20 for Test B. These sample means were fairly close to those of the means for the participating schools. The eighth grade means in the linkage sample were also similar to the means of the participating schools, approximately 60 and 20, respectively for the two tests. Both tests were reliable measures: For both grades, lower-bound reliability estimates for Test A were 0.90, and 0.80 for Test B.

Estimates of regression coefficients and the linkage error for both the full and final models appear in Table 5 for grade four, and Table 6 for grade eight. The estimates for the coefficients of the Test B school mean ($\bar{x}_{t..2}$), and of school proportions of students by minority ($\bar{d}_{t..1}$) and gender ($\bar{d}_{t..2}$) were not statistically significant. While the estimate for the coefficients of the Test A school mean ($\bar{x}_{t..1}$) in both grades was not statistically significant in the full model, it became statistically significant when the school proportions of students by minority and gender were removed from the model.

Figures 1a and 1b compare the actual and mean NAEP scores with respect to the score distributions of Test A (Figure 1a) or Test B (Figure 1b). Overall, the fits between the predicted and actual mean NAEP scores were pretty good. The model was less accurate where the frequency of students attaining a certain score on the state test was low. Figures 2a and 2b show similar results for the eighth grade tests. The estimates of error due to the linkage were quite large: 22 NAEP points for grade four, and 23 points for grade eight. It follows that the margin of error for a 95 percent confidence interval around an individual's predicted mean NAEP score would be 90 points on the NAEP scale. Given that the cutpoints were 30 (fourth grade) or 40 (eighth grade) points apart, the margin of error was too large to justify the predicting of individual scores from the model.

Tables 7 and 8 compare the classification of the NAEP achievement levels, based on the first plausible (hypothetical) value of each student in the sample, with the classifications from the predicted model[3]. For both grades, agreement was the lowest for the proficient level, was highest for the advanced level, and was consistent with the fact that in both grades the most and least numbers of students were classified, respectively, as proficient and advanced. Taken as a whole, Tables 9 and 10 suggest that the probability of incorrectly classifying a student, depending on grade and achievement levels, varies between 7 and 34 percent. This degree of uncertainty, again, is partially due to the use of the linkage to predict individual performance, and partially due to the inherent unreliability of NAEP as a measure of individual performance.

Figure 3 shows the estimated proficiency distributions for the fourth grade students in the sample. The three vertical bars represent the cutpoints between each proficiency level. For any

[3] The proficiency distributions in Tables 11 and 12 are different from those in Table 6. The former is based on the first plausible value for each student in the validation sample, whereas the latter is based on the mean plausible value of all students in the linkage sample.

projected NAEP score, the probability that a student with that score could have obtained a mean plausible value sufficiently high enough to be classified in each of the proficiency categories can be determined. Figure 4 shows the same information for eighth grade students. Possible interpretations from these distributions include the following:

1. A fourth-grade student with a projected NAEP score of 238 (the cutpoint between the basic and proficient levels) would have had a 50 percent chance of being classified at or above proficient.

2. An eighth-grade student with a projected NAEP score of 253 (10 points above the cutpoint for the basic level of proficiency) would have had a 67 percent chance of being classified at or above the basic level.

3. An eighth-grade student with a projected NAEP score of 293 (30 points below the cutpoint for advanced proficiency) would have had a 10 percent chance of being classified as advanced.

Estimating Achievement for Groups of Students

The variances due to sampling error in the linkage sample and due to linkage error are, respectively 1.73 and 23.9 for grade four, and 1.51 and 21.5 for grade eight. Suppose one wished to estimate the standard error for one group of fourth and eighth grade students, each numbering 10,000 students from 100 schools. The effective sample size, based on Equation (9), would be

918. Using Equation (10), one would obtain respective standard errors of 3.6 and 2.8. As a result, margins of error, assuming a 95-percent confidence interval, for the fourth- and eighth-grade groups mentioned above would be 7.5 and 8.5 points on the NAEP scale. One sees in Equation (10) that $\hat{\sigma}_p$, the standard error of the projected mean, is inversely proportional to the effective sample size, and has a minimum (asymptotic) value of $\hat{\sigma}_s$.

Discussion

In light of the *Goals 2000: Educate America Act*, which mandates that states monitor student achievement with respect to national standards, an increasing number of states consider NAEP to be the benchmark to which they ought to compare results (Williams, Rosa, McLeod, Thissen & Sanford, 1999). While the desire for using linkages is growing, some problematic issues inherent in the proper use of linkages merit consideration.

One of the requirements of a proper linkage is that the state assessment be similar in content to NAEP. An exaggeration that illustrates this point is the attempt to link a state's reading assessment to the NAEP mathematics assessment. Such a linkage is mathematically possible, though it might not be of any value, since reading and mathematics are different constructs. At the other extreme is the attempt to link two parallel assessments. Such a linkage could be accomplished through linear or equipercentile equating methods (Kolen & Brennan, 1995), thus making projection unnecessary.

Typically, however, the state assessments that are to be linked to NAEP fall in between these two extremes. That is, while the two assessments are not likely to be parallel, it is not clear how extreme they differ in content. An item-by-item content analysis might prove useful in this

case. However, while such an analysis might be possible for the state assessment, it is not possible for NAEP, for only certain NAEP items released from previous administrations would be available.

A good state-NAEP linkage also requires that the state assessment be reliable (internally consistent) and that it correlate highly with NAEP. A high measure of reliability offers some assurance that a student retaking the state assessment under similar conditions (and assuming that the student doesn't learn any test-relevant information from one administration to the next) would receive a similar score. That is to say, the predictor measure, upon which the linkage is based needs to be reliable. A high correlation between the state assessment and NAEP reduces the amount of random error introduced into the linkage.

Standard errors for the regression coefficients and for the linkage were estimated through repeated half-sampling. Other methods are available, such as the bootstrap and jackknife. The repeated half-sampling method was used because it worked, and not out of theoretical considerations. Williams, Rosa, McLeod, Thissen and Sanford (1999) used the bootstrap to estimate standard errors on the grounds that the jackknife may yield biased estimates, possibly due to within-school variability. A comparison between repeated half-sampling, bootstrap and jackknife methods of error estimation is warranted, and may provide some revealing results.

Just as NAEP is designed to measure achievement for *groups* of students, the linkage in this study is not intended to measure the achievement of individual students. With the linkage error of individual scores around 20 points on the NAEP scale, the margin of error for projecting a student's expected NAEP score is between 80 and 90 points, and is large enough to classify a student who is performing below basic achievement as advanced. The linkage error is inversely proportional to the effective sample size, which is a function of the size of the sample to which

the linkage is being applied. Thus, larger sample sizes for the target sample will result in smaller linkage errors.

It is easy to ignore the error in the linkage, especially from the perspective of test publishers. However, it is *impossible* at present to project individual NAEP scores from *any* state assessment with a sufficient accuracy that would warrant reporting individual scores.

Conclusion

As more and more states turn towards NAEP as a benchmark for student achievement, there will be a growing need to develop accurate linkages. The state-NAEP linkage proposed in this study demonstrates that state assessments can be linked to national standards, although this methodology deserves more study.

References

Allen, N. L. & Johnson, E. G. (1996). The design and implementation of the 1994 NAEP. In Allen, N. L., Klein, D. L., & Zelenak, C. A., *The NAEP 1994 technical report*. (NCES Publication No. 97-897). Washington, D.C.: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Allen, N. L., Johnson, E. G., Mislevy, R. J., & Thomas, N. (1996). Scaling procedures. In Allen, N. L., Klein, D. L., & Zelenak, C. A., *The NAEP 1994 technical report*. (NCES Publication No. 97-897). Washington, D.C.: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Ballator, N., Jerry, L. & Rogers, A. (1999). *NAEP 1998: State reading report for Connecticut*. (NCES Publication No. 1999-460CT). Washington, D.C.: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating*. New York: Springer-Verlag.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McLaughlin, D. H. & Arenson, E. A. (1999). *NAEP-state reading assessment linkage study*. Unpublished manuscript.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Wallace, L. & Rust, K. F. (1996). Sample design. In Allen, N. L., Klein, D. L., & Zelenak, C. A., *The NAEP 1994 technical report.* (NCES Publication No. 97-897). Washington, D.C.: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1999). Projecting to the NAEP scale: Results from the North Carolina End-of-Grade testing program. *Journal of Educational Measurement, 35,* 277-296.

Table 1. Position of blocks within booklets in the 1998 NAEP reading assessment, grade 4.

	Block							
Booklet	C	D	E	F	G	H	I	J
1	second	first						
2	first		second					
3			first				second	
4		second					first	
5		first	second					
6	first						second	
7				first				second
8					second			first
9					first	second		
10				second		first		
11				first	second			
12						second		first
13		second			first			
14	second					first		
15			first	second				
16							first	second

Table 2. Position of blocks within booklets in the 1998 NAEP reading assessment, grade 8.

	Block									
Booklet	C	D	E	F	G	H	I	J	K	M
1	first	second								
2		first	second							
3	second		first							
4				first		second				
5					second	first				
6				second	first					
7							second	First		
8							first		second	
9								Second	first	
10	first					second				
11		second			first					
12			first	second						
13				first			second			
14						first			second	
15					second			First		
16		first								
17			second				first			
18	second								first	
21										*

* Booklet 21 consisted of one 50-minute block. All other booklets contained two 25-minute blocks.